Possibility of large deviations in optimization algorithms

Boris Polyak

(Institute of Control Science, Moscow, Russia)

with P. Scherbakov, M.Danilova, A.Kulakova

October 27, 2017, MPhTI

OUTLINE

- Polynomial-time, O, O^* – be careful!
- $O(\mathrm{e}^{-\lambda t})$, $O(q^k)$ and real behavior – examples
- Rigorous results on linear difference equations
- Behavior of heavy-ball and Nesterov's accelerated gradient algorithms

How good are polynomial-time algorithms?

- Dyer, Frieze, Kannan A random polynomial-time algorithm for approximating the volume of convex bodies 1991

$$N = O^*(n^{23}) \leq 10^{15} n^{19} \frac{1}{\varepsilon}$$

  Lovasz, Vempala 2006

$$N \leq 10^{10} n^3 \log \frac{1}{\varepsilon}$$

  Cousins, Vempala 2016 $n = 100$ computationally tractable

- Ellipsoids method is polynomial-time for convex optimization...

Estimates $O(\mathrm{e}^{-\lambda t})$

Known facts: $\dot{x} = Ax, \quad x \in \mathbb{R}^n,$

for Hurwitz stable $\Re\lambda_i(A) \le -\lambda < 0$ have $x(t) = O(\mathrm{e}^{-\lambda t})$

BUT: Moler, Van Loan, 19 dubious ways to compute the exponential of a matrix, 25 years later, *SIAM Review*, 2003.

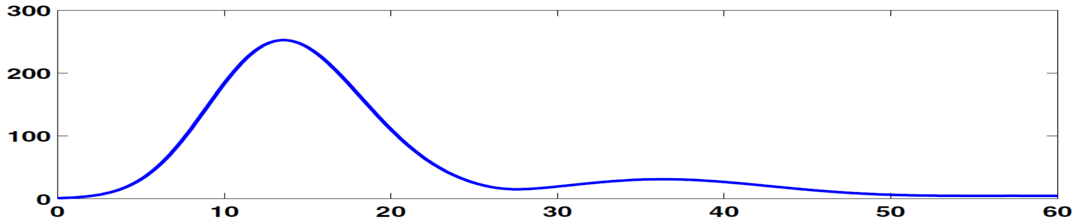


Fig. 3 $\|e^{tA}\|$, *the hump for the transient example.*

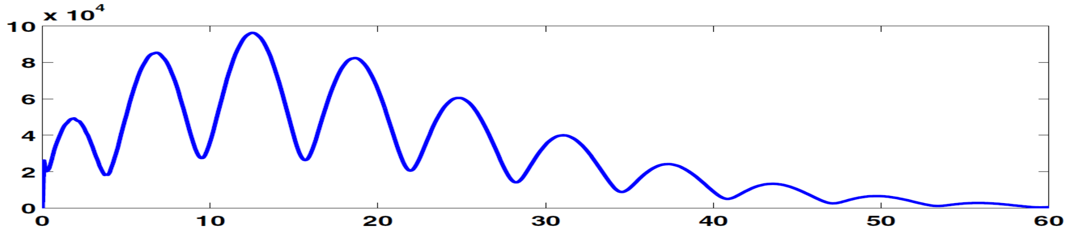


Fig. 5 $\|e^{tA}\|$, *the hump for the stabilized Boeing 767.*

Estimates $O(\mathrm{e}^{-\lambda t})$ Contd

Godunov 2002 — examples with huge peaks for Hurwitz systems with two-diagonal matrices.

Polyak, Smirnov *Automatica 2016* Rigorous results on *lower* bounds for deviations in stable systems with matrices in companion form.

Estimates $O(q^k)$

$$x_{k+1} = Ax_k$$

A is Schur stable: $\rho = \max_i |\lambda_i(A)| < 1, ||x_k|| = O(\rho^k)$.

Similarly: for scalar difference equations

$$x_{k+1} = a_1 x_k + a_2 x_{k-1} + \cdots + a_n x_{k-n+1},$$

$|x_k| = O(\rho^k), \rho = \max_i |\rho_i|, \rho_i$ being the roots of characteristic polynomial

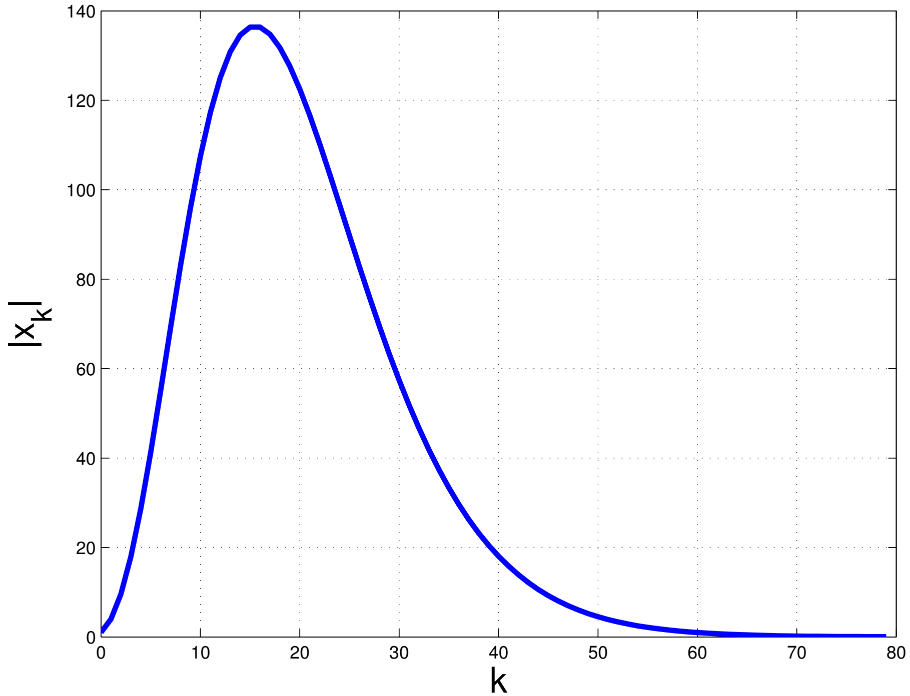
$$p(s) = s^n - a_1 s^{n-1} + \cdots - a_n$$

Thus for $\rho < 1$ $x_k \to 0$.

Example

Let $p(s) = (s - \rho)^n, x_1 = x_2 = \cdots = x_{n-1} = 0, x_n = 1$. Then

$$x_k = C_{k-1}^{k-n} \rho^{k-n}$$



Peak of trajectory: $n = 5$, $\rho = 0.8$, $x_{\text{init}} = (0, \ldots, 0, 1)$

Example

n	2	3	4	5	6	7
x^*	1	2.963	16.519	136.375	$1.494 \cdot 10^3$	$2.041 \cdot 10^4$
k^*	3	6	12	20	30	42

Dependence of x^* and k^* on n for $\rho = 1 - \frac{1}{n}$ and $x_{\text{init}} = (0, \dots, 0, 1)$

Lower bound for difference equations

Theorem 1 If ρ_i are all real and $0 < \rho \le \rho_i < 1$, then there exist initial conditions $|x_i| \le 1, i = 1, \ldots, n$ such that $|x_k| \ge C_{k-1}^{k-n} \rho^{k-n}$. In particular $|x_k| \ge (k-1)\rho^{k-2}$ for $n = 2$.

Thus "peak effects" are unavoidable for stable difference equations with real roots of the characteristic polynomial, close to 1.

Matrix difference equations

Let $x_i \in \mathbb{R}^m$, be generated by difference equation

$$x_{k+1} = A_1 x_k + A_2 x_{k-1} + \cdots + A_n x_{k-n+1}, A_i \in \mathbb{R}^{m\times m}$$

with initial condition $x_1, \ldots, x_n$.

Theorem 2 If matrices A_i commute and roots ρ_i of all polynomials

$$p(s) = s^n - a_1 s^{n-1} + \cdots - a_n, a_i \in \sigma(A_i)$$

satisfy $|\rho_i| \leq \rho < 1$ then

$$||x_k|| = O(\rho^k)$$

Lower bound for matrix difference equations

Theorem 3 If under above conditions ρ_i are all real and $0 < \rho \leq \rho_i < 1$, then there exist initial conditions $||x_i|| \leq 1, i = 1, \ldots, n$ such that $||x_k|| \geq C_{k-1}^{k-n} \rho^{k-n}$. In particular $||x_k|| \geq (k-1)\rho^{k-2}$ for $n = 2$.

Optimization

We consider L-smooth, l-strongly convex unconstrained minimization in Euclidean space

$$\min f(x), x \in R^n,$$

$$||f'(x) - f'(y)|| \leq L||x - y||, f(x + y) - f(x) - (f'(x), y) \geq \frac{l}{2}||y||^2$$

Denote $\kappa = \frac{L}{l}$ its condition number, $x^* = \arg\min f(x), f^* = \min f(x)$.

Then gradient method

$$x_{k+1} = x_k - \alpha f'(x_k)$$

converges monotonically in x and f: $||x_k - x^*|| \leq ||x_0 - x^*||q^k$. The best $q = \dfrac{L - l}{L + l}$ is for $\alpha = \frac{2}{L+l}$. For κ large q is close to 1: $q \approx 1 - 2/\kappa$.

Faster algorithms

First-order algoritms:

- Conjugate gradient
- Heavy ball
- Nesterov's accelerated gradient
- "The fastest known algoritm"

All of them have the same complexity as gradient method, don't exploit matrices, well suited for large-dimensional problems and found wide application in deep learning.

Stationary versions

- Heavy ball *HB* (Polyak 1964)

$$x_{k+1} = x_k - \alpha f'(x_k) + \beta(x_k - x_{k-1})$$

- Nesterov's accelerated gradient *NA* (Nesterov 1983, Nesterov book, (2.38))

$$x_{k+1} = y_k - \alpha f'(y_k)$$

$$y_{k+1} = x_{k+1} + \beta(x_{k+1} - x_k)$$

- "The fastest known algorithm" *FK* (Van Scoy, Freeman, Lynch 2018)

All of them have the same complexity as gradient method, don't exploit matrices, well suited for large-dimensional problems and found wide application in deep learning.

Quadratic case

$$f(x) = \frac{1}{2}(Ax, x), lI \leq A \leq LI, x^* = 0, f^* = 0, f'(x) = Ax$$

Best parameters and initial values, known results:

- *HB* $\alpha = \frac{4}{\sqrt{L}+\sqrt{l}}, \beta = q^2, q = \frac{\sqrt{L}-\sqrt{l}}{\sqrt{L}+\sqrt{l}}, x_0 = x_1$

  Then $||x_k|| = O(q^k)$. More precise estimate?

- *NA* $\alpha = \frac{1}{L}, \beta = q, x_0 = y_0,$ Then $f(x_k) \leq (f(x_0) + l/2||x_0||)(1 - \frac{1}{\sqrt{\kappa}})^k$

- *FK* $\alpha^*, \beta^*, \gamma^*, \delta^*$ Then $||x_k|| \leq \sqrt{\kappa}||x_0||(1 - \frac{1}{\sqrt{\kappa}})^k$

Comparison: power terms and pre-power terms.

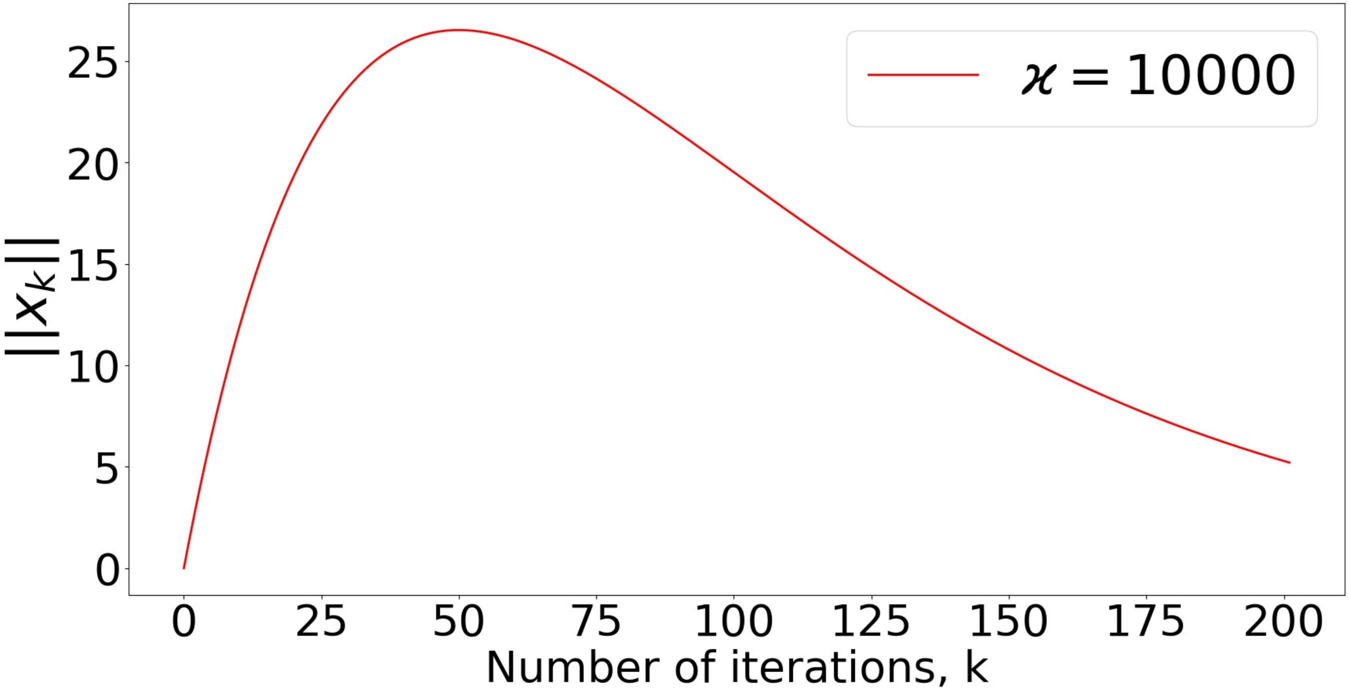
Refining estimates for HB

$x_{k+1} = ((1+\beta)I - \alpha A)x_k - \beta x_{k-1}$ Characteristic polynomials in Theorem 2 above are $p(s) = s^2 - (1+\beta-\alpha\lambda_i)s + \beta, l \le \lambda_i \le L$. For $\alpha = \alpha^*, \beta = \beta^*, \lambda_i = l$ the polynomial has two equal roots q. Thus if $x_0 = 0, x_1 = h, Ah = lh$ then for κ large $||x_k||$ monotonically grows until $k = \sqrt{\kappa}$ achieving the value $\max_k ||x_k|| \approx \sqrt{\kappa}/e$

However such initial conditions do not arise for standard version of HB when $x_0 = x_1$.

Numerical tests

HB, $A = \mathrm{diag}\,(1, 10000), x_0 = (0, 0)^T, x_1 = (1, 1)^T, \alpha^*, \beta^*$



$f(x_k)$ behaves in similar way. If $x_0 = x_1$ peak effects are lacking.

Analysis for NA

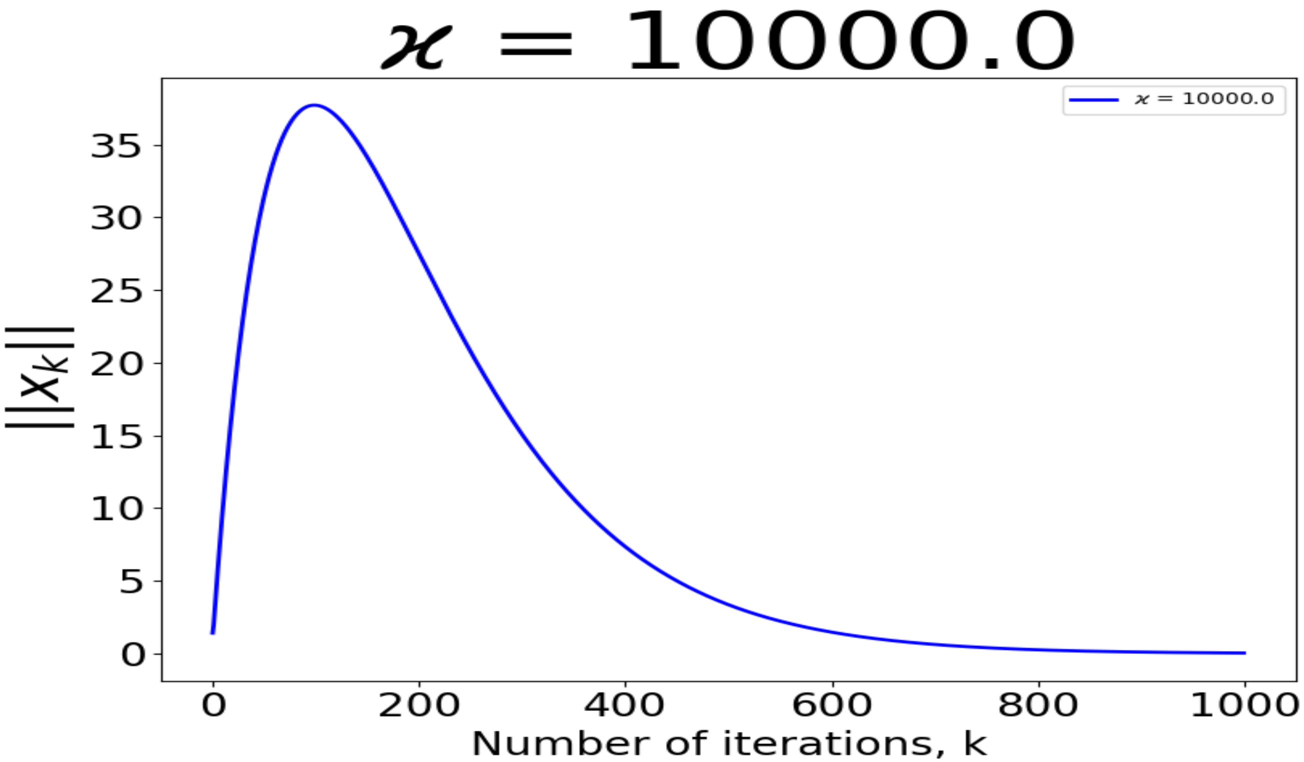
Eliminating y:

$x_{k+1} = (I - \alpha(1+\beta)A)x_k - \alpha\beta A x_{k-1}$ Characteristic polynomials in Theorem 2 above are $p(s) = s^2 - ((1 - \alpha(1+\beta)\lambda_i)s + \alpha\beta\lambda_i, l \leq \lambda_i \leq L$. For $\alpha = \alpha^*, \beta = \beta^*$, analysis can be performed based on Theorem 3 and we conclude that for some initial conditions

However such initial conditions do not arise for standard version of NA when $x_0 = y_0$.

Numerical tests

NA, $A = \operatorname{diag}(1, 10000), x_0 = (0, 0)^T, y_0 = (1, 1)^T, \alpha^*, \beta^*$



$f(x_k)$ behaves in similar way. If $x_0 = y_0$ peak effects are lacking.

Upper bounds

We can get nonasymptotic estimates by use of quadratic Lyapunov function. If we write our method (HB or NA) in the form

$$z_{k+1} = Hz_k, z_k = (x_k, x_{k-1})^T$$

and construct Lyapunov function $V(z) = (Pz, z)$ by solving Lyapunov equation

$$H^T PH - P = -Q, Q > 0$$

then we obtain estimates like

$$||z_k|| \leq C||z_0||r^k$$

with r, C depending explicitly on P, Q. By choosing Q in some optimal manner we get trade-off between r, C. This is a direction for future research.

Future work

- Upper bounds - explicit results
- Simulation for hard quadratic functions
- Nonquadratic functions
- Algorithms with errors in gradient